

Knowledge Bases for Computerized Physical Property Estimation

Kevin G. Joback

Molecular Knowledge Systems, Inc.

PO Box 10755, Bedford, NH 03110-0755

Keywords

Model, molecular simulation, property estimation, group contributions, databases

Abstract

The wide availability of computer software for statistical analysis, chemical structure manipulation, and process simulation has resulted in the development of thousands of physical property estimation techniques. Many of these techniques are highly tailored, applicable only to a specific chemical family or a narrow range of state conditions. Computerized management of estimation techniques is thus essential today. We discuss a new class of databases, called knowledge bases, that, in addition to manipulating numeric and textual physical property data, manipulate molecular structures and physical property estimation techniques.

Introduction

The majority of estimation techniques used today can be classified as either group contribution or equation oriented. From a computerization perspective equation oriented techniques are a special case of group contribution techniques, i.e., group contribution techniques without any groups. Three core steps are needed to estimate a chemical's physical property using a group contribution technique:

1. Select an accurate and applicable technique by considering the chemical's molecular structure and the given state conditions.
2. Determine the number of occurrences of each technique's group within the chemical's molecular structure.
3. Collect and input group contributions into the technique's model to generate a final estimate.

We discuss how we implemented these steps in a physical property estimation software package called Cranium. Cranium is a complete database capable of managing over 50 different physical properties. Cranium's knowledge base capabilities enable it to store and manipulate molecular structures and estimation techniques in the same manner traditional databases store and manipulate text and numeric data.

Selecting Accurate Estimation Techniques

Cranium's knowledge bases store data in an object-oriented framework. Each boiling point, melting point, molecular structure, etc., is represented as an object called a "datum". Each

datum stores a property value, associated state conditions (temperature, pressure, composition, etc.), reference information, textual comments, and accuracy limits. The equations used by estimation techniques are also stored as objects and manipulated just like any other piece of data. Unlike the integrally encoded equations found in most process simulators, knowledge bases enable models to be added, removed, edited, copied and pasted just like any other type of data.

The datum used to manage a technique's estimation model stores both a text version for easy human manipulation and a compiled version for rapid computer manipulation. Whenever the text version is changed the computer system generates a compiled version, which also verifies the correctness of the entered code, and stores these instructions back into the data object. This object-oriented basis enables the easy creation and management of knowledge bases containing hundreds of estimation techniques.

In addition to storing each technique's estimation model, Cranium's knowledge bases also store another datum representing each technique's "preamble". The preamble contains code that examines a molecule's structure or a mixture's components and determines whether or not the technique is applicable. For example, some estimation techniques should not be used for polar compounds or for mixtures that contain water. The preamble code captures this knowledge for each technique. If the technique is found to be applicable, the preamble also returns a number indicating the technique's accuracy.

The preamble code is given the same arguments as the technique's model code. These arguments are typically a reference to the material or mixture whose properties are currently being estimated and state variables such as temperature, pressure, and composition. The preamble can perform very simple evaluations such as ensuring the given mixture is a binary or more complex evaluations such as ensuring the entered temperature is less than 0.80 times the critical temperature.

Chemical family membership is often used to organize the applicability of estimation techniques. For example, Brock and Bird's surface tension technique [1] is not applicable to polar compounds and Klincewicz's critical temperature technique [2] should not be used for heavily fluorinated compounds. Many mixture estimation techniques are specific for aqueous or hydrocarbon solutions. Chemical family membership is also often used to organize technique accuracy. For example, the Macleod and Sugden surface tension technique yields an average absolute percent error of 12% on ketones, 14% on amines, and 24% on esters [3].

Although any generalization of estimation technique accuracy risks oversimplification, these chemical family averaged errors represent valuable knowledge that should be utilized. Cranium's knowledge bases thus record chemical family membership for each stored material and mixture. Table 1 shows the list of currently available chemical families. A technique's preamble code can then check family membership in a series of if-then statements assigning the appropriate accuracy. Figure 1 shows such an example code fragment.

In some cases the granularity of information offered by chemical family classification is too coarse. In these cases advantage is taken of Cranium's ability to analyze molecular structure in detail. (These abilities are detailed later in this paper.) For example, certain estimation techniques may only be applicable to low molecular weight polyols. Polyol is not an available chemical family but each technique's preamble code can execute functions that perform substructure searches.

Cranium's overall estimation procedure thus begins by collecting a list of all techniques that are capable of estimating the desired physical property. Each of these techniques' preamble code is then executed. Techniques that are not applicable are removed from the list. The applicable techniques are then sorted by their accuracy value. Cranium then chooses the most accurate technique and executes its model code. If the model code returns false, for example if the molecule's structure can not be represented by the technique's groups, then the next most accurate technique is tried. In this manner the most accurate of the applicable estimation techniques will always be used to provide an estimate.

Determining Group Occurrences

Once a group contribution technique has been selected the occurrence of each of the technique's groups within the molecule's structure must be determined. Numerous substructure search algorithms have been reported in the literature [4]. Cranium's implementation uses a network-matching algorithm.

In this algorithm we represent structures, atoms and bonds as objects. An atom records information on its display coordinates, its chemical element, and a list of attached bond objects. Each bond object stores information on its type and its two attached atom objects. This cross exchange of atom and bond references is what establishes the network. It is thus possible to start at any atom or bond and traverse the entire structure.

The substructure groups that comprise the estimation technique are also represented as atoms and bonds. However, most groups contain a special class of atom called a “free” atom. Free atoms represent “wildcards” and are used for matching. Table 2 shows several example groups. The asterisk denotes a free atom. The free atom’s enclosing square brackets denotes that it is “nonsubtractable”.

The concept of atom subtractability is used to represent the nearest neighbor and next-nearest neighbor restrictions used by second order group contribution techniques such as Benson’s [5]. Table 2 shows several groups with nonsubtractable atoms. Group A is a simple methyl group which represents a carbon atom bonded to three hydrogen atoms and any other atom. The free atom places no restriction on what this neighbor must be. Group B is a methyl group bonded to a nonsubtractable nitrogen. In this group the carbon atom must be bonded to three hydrogens and single bonded to any type of nitrogen. Group C further requires the bonded nitrogen to be an imine.

Some estimation techniques distinguish between acyclic and cyclic groups. For example, in Lydersen’s critical temperature technique [6] the -CH₂- group has very different

contributions depending upon whether or not it is in a ringed structure. However other techniques, e.g., UNIFAC [7], do not make any distinction. A -CH₂- group in UNIFAC has the same contributions whether it is in a ring or not.

Cranium's knowledge bases require the type of each bond be explicitly specified. In addition to acyclic bond types, Cranium has explicit cyclic and aromatic bond types. Table 3 shows the graphical notation used to represent an acyclic -CH= group, a cyclic -CH= group, and an aromatic CH group.

The explicit enumeration of bond types does increase the number of groups. Table 4 shows the four groups that result from a >C< group when its bonds' types are explicitly specified.

An estimation technique's groups often represent redundant structural information. This redundancy will result in multiple ways in which a molecular structure can be dissected into groups. Table 5 shows two possible dissections of acetic acid into groups from a hypothetical estimation technique. Larger molecules may have many more possible dissections [8].

Much of this redundancy can be explained by the need to represent more specific structural information. The proximity of the >C=O and -OH groups within a -COOH group causes different chemical behavior. Hence a new, more specific group, must be introduced to represent this unique behavior. Some estimation technique developers [9] have found that

specific groups containing four or more “fundamental” groups are needed to provide adequate accuracy.

Many techniques use “corrections” that would be better implemented as “specific” groups. For example Domalski and Hearing’s extension of Benson’s technique [10] has contributions for -CH₂- and >CH- groups and a correction value for a cyclobutane ring. Table 6 shows several group contributions and corrections for Benson’s enthalpy of formation technique [10] and illustrates how these groups and contributions can be combined into a new set of specific groups and contributions. Although the number of groups increases using this approach any ambiguity associated with the use of a group is eliminated.

Cranium sorts all groups before dissection to ensure specific groups are used before fundamental groups. Each group is assigned a “specificity index” according to Equation 1:

$$\text{Index} = 10000 * \#(\text{subtractable atoms}) + 100 * \#(\text{nonsubtractable atoms}) + \#(\text{free atoms}) \quad (1)$$

Thus a >C=O group would have an index of 20002, an -OH group would have an index of 20001, and a -COOH group would have an index of 40001. Of these three groups the -COOH is the most specific and hence would be tried first.

Ultimately the correct order of groups for dissection is the one that matches the order used when the estimation technique was developed. Cranium’s knowledge bases thus enable

specificity indices to be manually entered for any group. These manually entered indices will be used instead of the index calculated by Equation 1.

Once the technique's groups have been sorted, dissection begins by examining the most specific group first. Each group's atoms are matched in a depth-first manner against the atoms contained in the structure being estimated. Two atoms match when they have the same chemical element and are connected by the same types of bonds to the same types of neighbors. Free atoms are not required to match chemical elements and must only match a subset of neighboring atoms. Other nonsubtractable atoms must match elements but also only need to match a subset of neighboring atoms. Once a match has been made the matching group's atoms are subtracted from the structure. The term subtraction simply implies that the matched structure atoms are no longer available for further matching.

Nonsubtractable atoms are used to restrict the matching of groups. However, they are not subtracted when a group is subtracted and thus must be further matched to other groups for a correct dissection. For example, although the CH₃-[N] group occurs once in methyl amine, its contribution must be combined with that of the -NH₂ group for a correct estimation. Figure 2 shows how the concept of atom subtractability addresses this procedure. Once a group is found to match a portion of a molecular structure only its subtractable atoms are subtracted from the structure. The remaining atoms, including those that matched nonsubtractable group atoms, are available for further matching.

Collecting Group Contributions

Once a molecule's structure has been dissected into a set of group occurrences the technique's contributions must be obtained. Cranium's knowledge bases store contributions in a keyword-indexed table – given a unique keyword a numeric contribution is retrieved. For many estimation techniques the keyword is simply the name of the group. For estimation techniques that use binary group interactions, e.g. UNIFAC [7] the keyword is formed by appending the names of two groups. Keywords would be similarly formed for techniques that have ternary or higher group interactions.

Some estimation techniques have several contributions for each group. For example in Joback's ideal gas heat capacity technique [11] each group has a different contribution for each of four polynomial coefficients. These contributions are also stored in keyword-indexed tables. However, a second term is added to each keyword to indicate which coefficient the contribution is for.

Storing contributions in keyword-indexed tables provides a simple and consistent representation. It is also easier to add new groups to a keyword-indexed table than it is to expand the vectors and arrays typically used for group contribution storage.

Conclusion

The continuing goal of our research is to develop a computer software program that can estimate all the properties of all types of materials using all kinds of estimation techniques. Representing the variety of models, groups, and contributions used by estimation

techniques has been one of our major challenges. Knowledge of the applicability and accuracy of each technique is also considered critical information that needs to be recorded. We believe we have addressed these challenges by using object oriented databases, preamble code to screen and rank techniques, automatic dissections of molecular structures into groups, free atoms and nonsubtractable atoms to increase group expressiveness, and keyword-indexed table lookup to systematize the storage and retrieval of contributions. Further research and development will continue to test the limitation of our approach.

References

- 1) J. R. Brock and R. B. Bird. "Surface Tension and Principle of Corresponding States". AIChE Journal, volume 1, page 174, 1955.
- 2) K. M. Klinecicz and R. C. Reid. AIChE Journal, volume 30, page 137, 1984.
- 3) Martin S. High, Manoj Nagvekar, Ronald P. Danner, and Thomas E. Daubert. "Documentation of the Basis for Selection of the Contents of Chapter 7 Surface Tension". AIChE, New York.
- 4) Neil A. B. Gray. "Computer-Assisted Structure Elucidation". John Wiley & Sons, Inc. 1986.
- 5) Sidney William Benson. "Thermochemical Kinetics". Wiley, New York. 1976.
- 6) A. L. Lydersen. "Estimation of Critical Properties of Organic Compounds". Univ. Wisconsin Coll. Eng., Eng. Exp. Stn. Rept. 3. Madison, Wisconsin. April 1955.
- 7) Aage Fredenslund. "Vapor-Liquid Equilibria Using UNIFAC: A Group Contribution Method". Elsevier, Amsterdam, 1977.

- 8) John W. Raymond and Tony N. Rogers. "Molecular Structure Disassembly Program (MOSDAP): A Chemical Information Model to Automate Structure-Based Physical Property Estimation". Journal of Chemical Information and Computer Sciences. September 1998.
- 9) Corwin Hansch and Albert Leo. "Exploring QSAR". American Chemical Society, Washington DC, 1995.
- 10) Eugene S. Domalski and Elizabeth D. Hearing. "Estimation of the Thermodynamic Properties of C-H-N-O-S-Halogen Compounds at 298.15 K". J. Phys. Chem. Ref. Data, volume 22, number 4, 1993, page 805.
- 11) Kevin G. Joback and Robert C. Reid. "Estimation of Pure-Component Properties from Group-Contributions". Chemical Engineering Communications, volume 57, page 233. 1987.

Figure Captions:

Figure 1: Example “Preamble” Code Fragment

Figure 2: Subtracting Groups from Molecular Structure

Table 1: Currently Available Chemical Families

Acid Bromide	Acid Chloride	Alcohol
Aldehyde	Amide	Amine
Aromatic	Binary	Brominated
Carboxylic Acid	Chlorinated	Contains H ₂ O
Contains H ₂ S	Ester	Ether
Fluorinated	Halogenated	Hydrocarbon
Iodinated	Ketone	Nitrile
Oxygenated	Perfluorocarbon	Phenol
Primary Amine	Saturated	Secondary Amine
Sulfide	Tertiary Amine	Thiol
Unsaturated		

Table 2: Example Groups with Free and Nonsubtractable Atoms

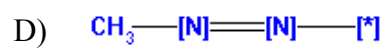
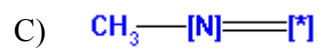
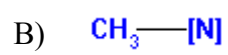
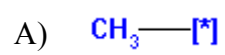


Table 3: Examples of Explicit Bond Types




Acyclic Group	
Cyclic Group	
Aromatic Group	

Table 4: Generated Specific Groups

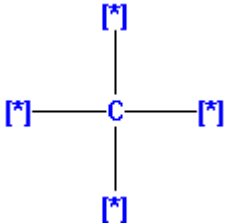
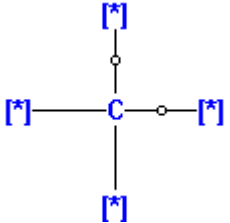
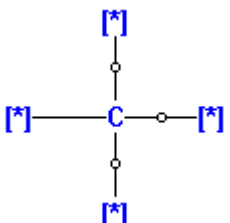
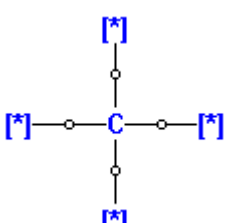
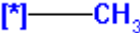
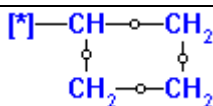

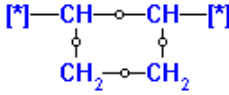
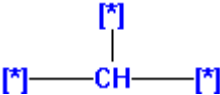
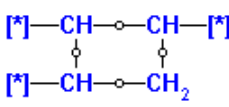
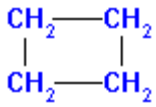
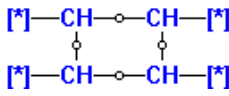
“Specific” Group	Example Structure
	Neopentane
	1,1-Dimethylcyclobutane
	Camphor
	Spiropentane

Table 5: Multiple Group Dissections of Acetic Acid

$\text{CH}_3-\overset{\text{O}}{\parallel}{\text{C}}-\text{OH}$	
Possible Group Dissection	Possible Group Dissection
$\text{CH}_3-[*]$	$\text{CH}_3-[*]$
$[*]-\overset{\text{O}}{\parallel}{\text{C}}-[*]$	$[*]-\overset{\text{O}}{\parallel}{\text{C}}-\text{OH}$
$\text{OH}-[*]$	

Table 6: Specific Groups accounting for “Corrections”

Original Group Contributions		“Specific” Group Contributions	
	-42.26		47.83
	-20.63		67.29
	-1.17		86.75
	110.89		106.21

```
// Retrieve chemical families
MatFamilies(material, families, ier);

// Assign technique accuracy
if( member("Ketone", families) )
    SetResult(12.0);

else if( member("Amine", families) )
    SetResult(14.0);

else if( member("Ester", families) )
    SetResult(24.0);
```

Figure 1: Example “Preamble” Code Fragment

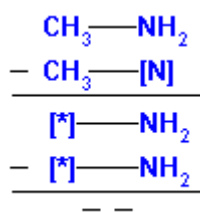


Figure 2: Subtracting Groups from Molecular Structure